第 41 卷 第 3 期
Vol. 41 No. 3
山 东 大 学 学 报 (理 学 版)
JOURNAL OF SHANDONG UNIVERSITY
2006 年 6 月
Jun. 2006

文章编号:1671-9352(2006)03-0090-05

基于本体语义的定题爬虫

郑健珍[1],林坤辉[1],周昌乐[2],康 恺[1]
(1. 厦门大学 软件学院,福建 厦门 361005;2. 厦门大学 信息科学与技术学院,福建 厦门 361005)

摘要:定题爬虫能迅速获取网络上特定主题的大量信息,对专业搜索引擎及数据挖掘应用都具有重大价值.针对目前通用的基于关键词主题过滤策略的不足,在概念聚集思想启发下,提出了基于本体语义的主题过滤策略.同时根据网页具有不同位置不同信息重要性的特点,提出了改进的加权特征项权值计算公式,实现基于语义的网页实时过滤.为进一步提高爬虫的工作效率提出链接相关度预测算法.对比实验表明此策略具有可行性.

关键词:定题爬虫;主题过滤;本体语义;链接分析
中图分类号:TP391.3 **文献标识码**:A

Ontology based on focused crawler

ZHENG Jian-zhen[1], LIN Kun-hui[1], ZHOU Chang-le[2] and KANG Kai[1]
(1. Software School, Xiamen Univ., Xiamen 361005, Fujian, China;
2. Information Science and Technique Dept., Xiamen Univ., Xiamen 361005, Fujian, China)

Abstract: Focused crawler can fetch large quantities of domain resources from the Web in a short time. It is very helpful in both foused search engines and data mining companies. In order to overcome the deficiency of topic filtering strategy based on key-words widly used nowadays, the paper proposed a topic filtering strategy based on concept elicited by concept congregation idea. The paper also proposed an authority modified weight calculation formula based on different importance of Web page information. By doing this, real time Web page filtering based on concept can be achieved. In the hope of improving focused crawler's work efficiency more, the paper also proposed a link forecast algorithm. At last, the comparative experiment shows that the strategies proposed in this paper are pratical.

Key words: focused-crawler; topic-filtering; ontology-semantic-analyse; hyperlink-analyse

0 引言

网络爬虫是因特网上一个自动下载网页的程序.网络爬虫已被广泛应用于搜索引擎.随着用户个性化与专业化需求的增加,传统爬虫已不能满足这种需求,因而出现了定题爬虫.定题爬虫会根据特定的抓取目标,有选择地访问网络链接,并迅速获取网络上特定主题的大量信息,因而对专业搜索引擎或需获取某主题信息进行数据挖掘的应用具有极大的价值.

为了实现特定领域信息的获取,需要某种主题过滤策略.目前通用的做法是根据网页中的关键词判定.但由于存在一词多义及一义多词的现象,这种基于关键词的判定策略已被证实精确度不高[1],会遗漏许多相关页面或添加许多噪音页面.因此我们提出一种基于语义的主题相关性判定策略,利用 ontology 对领域概念及概念间关系的明确定义来提高判定精度.

1 基于本体的定题爬虫框架

本文提出的框架结构由 2 个循环组成:网络爬

收稿日期:2006-03-29
基金项目:厦门大学 985 二期信息创新平台资助项目(0000-X07204)
作者简介:郑健珍(1982-),女,硕士研究生,主要研究方向为网络多媒体,智能信息处理.

行循环和本体循环(见图 1 中的 1 和 2). 网络爬行循环从页面抓取开始,按箭头的方向形成循环 1,不停地从网络中取得主题相关信息,将结果显示给用户,同时用户也可以修改控制参数控制页面的抓取. 其中进行主题过滤和链接分析时需要用到本体管理提供的本体知识作为评价依据. 另一个循环是本体循环,从本体管理开始到主题过滤,按箭头方向形成循环 2. 本体管理是用户先在领域专家的协助下明确本领域的共享概念及其概念间的关系,构建领域本体,然后根据在实际爬行过程中出现的高频率新概念进行本体的更新与维护.

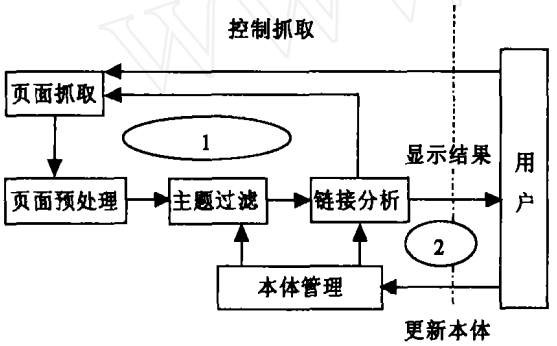

图 1　基于本体的定题爬虫框架

Fig. 1　Focused crawler framework based on ontology

2　基于本体语义的主题过滤

2.1　基于关键词的主题过滤

目前定题爬虫对网页相关性的判定仍然主要使用向量空间模型. 通常的做法是先从主题相关的示例网页集中提取关键词,作为本主题的特征项.

如果令 $p = \{p_1, p_2, \cdots p_n\}$ 表示网页集合,$k = \{k_1, k_2, \cdots k_t\}$ 表示主题特征集,k_i 为主题特征项,t 为特征项的个数,则网页 p_j 可表示成 $P_j = (w_{1,j}, w_{2,j}, \cdots w_{t,j})$,其中,$w_{i,j} = (\mathrm{tf}_{i,j} * \mathrm{idf}_j)$,$w_{i,j}$ 表示特征项 k_i 在网页 p_j 中的权值,$\mathrm{tf}_{i,j}$ 表示特征项 k_i 在网页 p_j 中出现的频率,idf_j 称为逆文献频率,为网页 p_j 中出现了特征项 k_i 的页面数的倒数.

而主题特征向量 R 可表示成 $R = (w_{1,r}, w_{2,r}, \cdots w_{t,r})$,$w_{i,r}$ 表示特征项 k_i 在主题特征向量 r 中的权值,可以在大量主题示例网页集中经过计算得出,也可以由专家给定.

因而网页 p_j 的主题相关性可由其对应向量 P_j 与主题特征向量 R 的夹角余弦来计算.

$$\mathrm{sim}(P_j, R) = \frac{P_j \cdot R}{|P_j| \times |R|} = \sum_{i=1}^{t} w_{i,j} \times w_{i,r} / \left(\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,r}^2}\right), \quad (1)$$

若 $\mathrm{sim}(P_j, R) \geq \theta$,$\theta$ 为设定阈值,则认为页面为主题相关,保存入数据库,否则丢弃.

这种算法的本质是统计页面中出现给定关键词的次数,出现次数越多,则认为越相关. 但实际情况是,一个概念可以用多个关键词表达,一个关键词也可表达多个概念,同时多个子概念表达的内容仍属于父概念的范畴,而这种基于关键词的相关性判定方法对这些情况都不能很好的处理. 看下面的例子,假设给定的主题特征项为"战斗机",当检索到这 2 个文档摘要时:

①歼击机,其特点是机动性好、速度快,空中战斗力强. 它们的首要任务是与敌机进行空战,夺取制空权;

② F-22 的空重为 13. 6 吨,最大起飞重量 27 吨,最大飞行速度 M2. 1,作战半径 1500 公里. F-22 的研制和生产总费用达到了 700 亿美元.

由于 2 个文档中都没有出现关键词"战斗机",因而用基于关键词计算得出的相关性比较低,很可能被抛弃.

观察这 2 个文档的关键词,①中的"歼击机",为"战斗机"的同义词,②中的"F-22",为战斗机的一种,属于"战斗机"的下义词. 因而这 2 个文档与关键词"战斗机"是语义相关的,属于要收集的内容. 如果能基于语义检索,则能收集到大量这种不出现关键词而语义相关的页面. 下面提出一种在概念聚集思想启发下提出的基于语义的主题相关性判定策略.

2.2　基于语义的主题过滤

要根据语义判定主题相关性,可采用从页面分析入手,将页面 p_j 中与主题词 k_i 具有相同概念的其他关键词(主要指其同义词和下义词)都替换成概念 k_i,这样页面 p_j 与主题特征向量 R 就能实现语义层次上的相似性判断,而且不会增加主题特征向量 R 的维数,也不存在新添加概念的权值确定问题.

由于 ontology 具有良好的概念层次及概念间、属性间关系定义,因而能很方便地获得一个词语的同义词或上、下义词. Hownet 是一个人工构建的目前已被广泛认可的 ontology,它以词语所代表的概念为描述对象,以揭示概念和概念之间以及概念所具有的属性之间的关系为基本内容[2]. 因此本文中直接采用 Hownet,而不另外构建 ontology. 一种做法是检查页面中所有名词(因为主题特征项一般都是名词)的同义词和 n 代上义词看能否与主题特征集中的特征项匹配,能的话则将该名词替换成该特征项,从而实现将关键词聚集成概念,如算法 1.

算法 1　页面语义化算法

<输入>页面 p_j 中每一个名词 s_i;主题特征集 k;基于语义的新页面 p_j'(初始时与原页面 p_j 相同);

<输出>每个特征项 k_i 在语义页面 p_j' 中总共出现的频率 $tf_{i,j}$;在页面中各个位置中出现的频率 $tf_{i,j,n}$,n 表示特征词出现位置(如标题,链接,加强文本等);

对每个 s_i {将 s_i 及用 Hownet 扩展出的 s_i 的同义词及 n 代上义词,存入集合 temp;

对 k 中的每个特征项 k_i

{If(k_i 在 s_i 的集合 temp 中)

{将 p_j' 中的 s_i 替换成 k_i;

$tf_{i,j}$ ++;//计算 k_i 出现频率

根据 s_i 在页面中位置 n,计算 k_i 在页面中不同位置出现的频率 $tf_{i,j,n}$;} } }

由于算法 1 需对页面中的每个名词都进行扩展,再进行匹配,会浪费许多时间在扩展主题无关名词上,使算法的效率不高,难于适应实时过滤的需求.实际上关键是要计算语义页面 p_j' 的特征向量 $P_j'=(w_{1,j},w_{2,j},\cdots w_{t,j})$.由于要实现网页的实时过滤,所以不采用传统的 TF-IDF 公式计算权值,而采用根据网页中信息的位置改进的权值计算公式.因而只需获得每个特征项 k_i 在语义页面 p_j' 中出现的频率 $tf_{i,j}$ 及 k_i 在语义页面中不同位置的出现频率 $tf_{i,j,n}$.根据这个需求,提出算法 2.算法 2 计算每个特征词的所有扩展词在页面 p_j 中出现的频率,也就是特征词在语义页面 p_j' 中的出现频率.这样达到了与算法 1 同样的效果,却大大降低了算法的复杂性,需要的只是简单的求和,适合实时过滤.

算法 2 页面语义化改进算法

对 k 中的每个特征项 k_i

{记集合 temp = { k_i, $k_{i,1}$, $k_{i,2}$, $\cdots k_{i,n}$ },其中 $k_{i,i}$ $(1\le i\le n)$ 表示用 Hownet 扩展出 k_i 的同义词或 n 代下义词;

将 temp 中各元素在页面 p_j 中出现的频率累加得 $tf_{i,j}$;

据 temp 中各元素在页面 p_j 中的位置得出 k_i 在页面中不同位置出现的频率 $tf_{i,j,n}$;}

网页上不同位置的信息具有不同的重要性(如标题文本一般比普通文本重要),我们提出了实时过滤的加权特征项权值计算公式

$$w_{i,j}=\lambda\times f_{i,j}\times w_{i,r}. \tag{2}$$

其中,$\lambda=\sum_{n=1}^{3}[f_{(n)}\times tf_{i,j,n}]$,$\lambda$ 表示位置加权系数,用特征词出现的位置权值 $f_{(n)}$ 与对应位置的频率 $tf_{i,j,n}$ 的乘积和来计算.$tf_{i,j,n}$ 可由算法 2 获得,而位置权值 $f_{(n)}$ 的确定,我们根据自己实践经验并参考别人的研究成果[3],认为网页中的锚文本(anchor text)最能反映页面内容,应赋予最高权值;而标题(title)、大标题(H1、H2)、加强文本(strong)也比较能反映页面内容,赋予次高权值.具体赋值如下:

$$f_{(n)}=\begin{cases}2, & n=1,\quad \text{在 anchor text 中}\\ 1.5, & n=2,\quad \text{在 title/H1/H2/strong 中}\\ 1, & n=3.\quad \text{其他}\end{cases}$$

$f_{i,j}=tf_{i,j}/\max_l tf_{l,j}$,表示页面 p_j 中的特征项 k_i 的标准化频率,即用特征项 k_i 在页面中出现的频率 $tf_{i,j}$ 除以在页面 p_j 中出现频率最高的特征项 k_l 的频率.$tf_{i,j}$,$\max_l tf_{l,j}$ 都可由算法 2 获得.

$w_{i,r}$ 表示特征项 k_i 预先给定的权值.

在计算出语义页面 p_j' 的特征向量 $P_j'=(w_{1,j},w_{2,j},\cdots w_{t,j})$ 后,就可以用公式(1)计算与主题特征向量 R 的相关性.

2.3 基于关键词和基于语义的主题过滤算法分析比较

为比较基于关键词和基于语义的主题过滤算法的性能,不失一般性设只有 2 个主题特征词 m,n,且 Hownet 中与 m 相似的词只有 m_1,m_2 2 个,与 n 相似的词只有 n_1,n_2 2 个.在页面 p_j 中 m,m_1,m_2;n,n_1,n_2 出现的标准化频率分别为 tf,tf_1,tf_2;tf′,tf_1',tf_2'.为了更具可比性,页面中特征词的权值计算统一采用公式(2).分别计算主题特征向量 R 与页面 p_j 的相关性.其中,[ ]表示向量,·表示内积,|| 表示模.则基于关键词的相关性为

$$\left.\begin{aligned}&R=[w_{m,r},w_{n,r}]\\ &P_j=[w_{m,j},w_{n,j}]=[\lambda_1\ tf\cdot w_{m,r},\lambda_2\ tf'\cdot w_{n,r}]\\ &\mathrm{sim}(R,P_j)=\frac{R\cdot P_j}{|R|\times|P_j|}=\frac{\lambda_1\ tf\cdot w_{m,r}^2+\lambda_2\ tf'\cdot w_{n,r}^2}{|R|\times|P_j|}\end{aligned}\right\}, \tag{3}$$

基于语义的相关性为

$$\left.\begin{aligned}&R=[w_{m,r},w_{n,r}]\\ &p_j'=[w'_{m,j},w'_{n,j}]=[\lambda_3\cdot(tf+tf_1+tf_2)\cdot w_{m,r},\\ &\qquad\lambda_4\cdot(tf'+tf_1'+tf_2')\cdot w_{n,r}]\\ &\mathrm{sim}(R,P_j')=\frac{R\cdot P_j'}{|R|\times|P_j'|}=\\ &\frac{\lambda_3\cdot(tf+tf_1+tf_2)\cdot w_{m,r}^2+\lambda_4\cdot(tf'+tf_1'+tf_2')\cdot w_{n,r}^2}{|R|\times|P_j'|}\end{aligned}\right\}. \tag{4}$$

比较公式(3),(4),从分母来看,虽然基于语义的新页面 p_j' 的模 $|P_j'|$ 会大于原页面 p_j 的模 $|P_j|$,但由于通常情况下页面向量的维数都已经非常大,

因此向量 P_j 与向量 P_j' 在长度上不会有太大差别,这种差别甚至可以忽略.因此影响相关性的主要是分子部分.由于 λ_3 是包括特征词 m,m_1,m_2 在内的位置加权系数,所以 λ_3 大于等于只包括特征词 m 的位置加权系数 λ_4.同时 $(tf + tf_1 + tf_2) \geq tf$,同理可得 $\lambda_4 \geq \lambda_2$,$(tf' + tf_1' + tf_2') \geq tf'$,因此基于语义的相关性判定会增强含有相关概念的页面的相关度,且性能较稳定.

3　链接分析对主题爬虫的改进

链接是爬虫工作的基础,主题相关页面中并非所有的链接都是主题相关的,在下载所有链接的页面内容前进行一次链接预测,去除那些明显不相关的链接,爬虫的效率能得到进一步提高,因为链接分析的复杂度远小于内容相关性判定的复杂度.所以,要进一步提高定题爬虫的工作效率,可从链接分析入手.

记 anchor_text_area < a href = "hyperlink" > anchor_text < /a > anchor_text_area,表示一个页面链接 hyperlink 的锚文本 anchor_text 及其链接附件文字 anchor_text_area.

分析在 anchor_text 及 anchor_text_area 中出现的主题特征词的数目,若超过一定阈值 γ,则预测 hyperlink 主题相关,放入待处理队列 Q 中等待下载页面内容.若低于阈值 γ,并不马上放弃,因为此 hyperlink 很可能是主题页面的前驱链接[4],也就是说可能目前看似不相关,但其后 N 层链接着相关内容.继续分析 anchor_text 及 anchor_text_area 中出现的前驱 N 层主题特征词的数目,若超过阈值 γ',则也给此 hyeperlink 继续被处理的机会,这样一定程度上能跳出定题爬虫局部搜索的通病.根据实践经验取 $N = 2$,前驱两层的主题特征集 k 由人工选择给出,anchor_text_area 取 30 个字节.

算法 3　链接相关度预测算法

<输入>:anchor_text,anchor_text_area,hyperlink,主题特征集 k,前驱 N 层主题特征集 k',阈值 γ,γ';

<输出>:待处理队列 Q;

对每一个 hyperlink

{检查 anchor_text 和 anchor_text_area 中出现主题特征集 k 中特征项的次数 f;

If $(f \geq \gamma)$ {将 hyperlink 放入待处理队列 Q;}

else{ 对每个前驱主题特征集 k'

{检查 anchor_text 和 anchor_text_area 中出现主题特征集 k 中特征项次数 f';

If $(f' \geq \gamma')$ { 将 hyperlink 放入待处理队列 Q;

return;//返回,当前 hyperlink 检查完毕,检查下一个 hyperlink}}

将 hyperlink 放入抛弃队列 Q';//hyperlink 即非主题相关链接也非前驱链接,抛弃 }}

4　实验与分析比较

4.1　实验设计

为验证本文提出的算法的性能,设计了 2 项测试.测试 1 比较基于关键词和基于语义主题相关性判定算法的查全率和查准率,引入新的性能判定指标[5]调和中数 $H(j) = \dfrac{2}{\dfrac{1}{r(j)} + \dfrac{1}{p(j)}}$,其中 $r(j)$ 和 $p(j)$ 分别表示前 j 个页面的查全率和查准率.只有查全率和查准率的值都较高时,调和中数 $H(j)$ 的值才会相应较高.测试 2 比较基于语义的主题相关性判定策略在加入链接预测前后在爬行速度和调和中数上的差别.

实验以"台湾军事情报"为特定主题,从《中国分类主题词表》中由人工精选得到主题特征集 k,并在一个约 500 k 的示例主题页面集中计算得出各主题特征项权值 $w_{i,r}$.同时根据这些示例页面集的前驱两层页面,提取出前驱 2 层主题特征集 k'.分词工具采用中文自动断词引擎开源项目 ZBNO.收集页面主题相关性的判定标准为:页面中出现在主题特征集中的短语数/页面中总的短语数 ≥阈值 η,经过示例主题页面集测试,该 η 取 0.3 较合适.初始 URL = 8(为以该主题在 google 上搜索得到的前 20 个 URL 人工精选后得出);搜索深度 = 3(避免时间过久,数据过多);同步线程数 = 50;主题相关性的设定域值 θ 取 0.3;链接相关度域值 γ,γ' 分别取 2 和 3.系统以 Java 语言实现,运行在一台装有 Windows Xp 系统,CPU 为 P4 1.5 G,硬盘为 160 G 的 PC 机上,通过 10/100 M 以太网卡接入因特网.

4.2　实验结果与分析比较

测试 1 的结果见图 2.可以看出基于语义的相关性判定算法比基于关键词的算法有更高的查全率和查准率,原因是本文提出的基于语义的主题过滤算法有其优点.测试 2 的结果见表 1.可以发现收集同样多页面,链接预测后比预测前所花时间少多了,

原因就是链接预测直接过滤掉了明显不相关的链接,节约了对这些链接进行较费时的基于内容的相关性判定时间,提高了爬虫的工作效率.但也发现链接预测后的调和中数有所下降,这是因为链接分析也有可能将个别相关链接视为不相关过滤掉.但比起工作效率的提高,这点精确度的下降是值得的,因而用链接分析改进爬虫工作效率也是可行的.

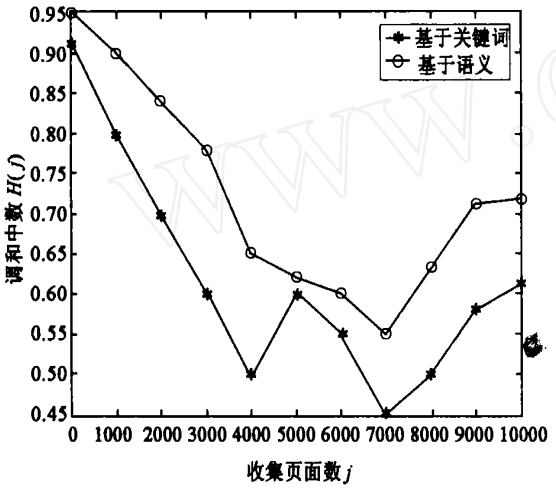

图 2 基于关键词和基于语义的性能比较

Fig. 2 Keywords based performance vs concept based performance

表 1 基于语义链接预测前后的性能比较

Tab. 1 Concept based performance before and after link forecasting

	收集页面数(个)	爬行时间(s)	查全率(%)	查准率(%)	调和中数 $H(j)$
链接预测前	3 000	930	74.7	81.6	0.780
链接预测后	3 000	720	70.5	73.1	0.718

上面的实验有较好的结果,是因为本文提出的算法与同类方法相比有如下的优点:

(1) 用概念聚集的思想实现语义层面上的主题过滤

不同于文献[6]扩充主题特征集语义,将每个主题特征项的同义词与下义词都添加进去,形成新的基于概念的主题特征向量的做法,本文的算法从页面入手,将页面中的同义词聚集成同一概念,在不改变主题特征集的情况下实现语义层面上的主题过滤.本做法不会出现文献[6]的向量维数急剧增加,加大算法复杂性,无法确定新添加概念的权值等问题,较好地实现了语义层面上的主题过滤.

(2) 用改进的加权特征项权值计算公式实现网页的实时过滤

不同于文献[7]下载完所有页面后再进行主题过滤的做法,本文算法结合网页的结构信息,提出了改进的加权特征项权值计算公式,实现了在爬行过程中的实时过滤,避免下载不相关页面,节省了存储空间,提高了爬虫的工作效率.

(3) 结合链接分析提高爬行效率

不同于传统的完全根据文本内容判断主题相关性,本文算法还添加了链接分析,在进行费时的内容相关性分析前先进行链接权值预测,大大提高了爬虫的爬行效率.而且在进行链接权值预测时,还考虑了链接成为 n 层前驱的可能性,一定程度上克服了定题爬虫局部搜索的局限性.

5 结语

基于本体语义的主题过滤策略比基于关键词的策略具有更高的判定准确性,同时进行链接预测能进一步提高爬虫工作效率.但本文所用到的本体语义还较有限,只涉及同义词及上下义词,可考虑结合领域本体中词语的属性及属性间关系,进一步提高主题过滤的准确性,这是下一步要做的工作.

参考文献:

[1] Marc Ehring, Alexander maedche. Ontology-focused crawling of Web documents[J]. Proceedings of the 2003 ACM Symposium on Applied Computing, 2003, 1(3):624～626.

[2] 董振东,董 强. Ontology 和 HowNet[EB/OL]. http://www.keenage.com/html/c_index.html., 2003-08/2006-02.

[3] Cutler M, Shih Y, Meng W. Using the structure of HTML documents to improve retrieval[A]. Proceedings of the USENIX Symposium on Internet Technologies and Systems Monterey[C]. California: California Press, 1997. 241～251.

[4] Mdiligenti F Coetzee. Focused crawling using context graphs[A]. Proceedings of the 26th International Conference on Very Large Data Bases[C]. Cairo: Cairo Press, 2000. 527～534.

[5] Ricardo Baeza-yates, Berthier Ribeiro-neto. Modern Information Retrieval[M]. Beijing: China Machine Press, 2005.

[6] 刘 林,汪 涛,樊孝忠. 主题爬虫的解决方案[J]. 华南理工大学学报,2004,32(11):137～141.

[7] 龙宇巍,王永成,许庆欢. 定题搜索引擎 Robot 的设计与算法[J]. 计算机仿真, 2004,21(4):70～76.

(编辑:孙培芹)